DOCUMENT RESUME

ED 464 949 TM 033 880

AUTHOR Williamson, David M.; Johnson, Matthew S.; Sinharay, Sandip; Bejar, Isaac I.

TITLE Applying Hierarchical Model Calibration to Automatically Generated Items.

INSTITUTION Educational Testing Service, Princeton, NJ.

PUB DATE 2002-04-00

NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Junior High School Students; Junior High Schools; Models; *Test Items

IDENTIFIERS *Calibration; Hierarchical Analysis

ABSTRACT

This study explored the application of hierarchical model calibration as a means of reducing, if not eliminating, the need for pretesting of automatically generated items from a common item model prior to operational use. Ultimately the successful development of automatic item generation (AIG) systems capable of producing items with highly similar statistical properties may permit the development and implementation of adaptive on-the-fly testing. The study applied the related siblings model to mathematics item data from an experimental administration associated with a national testing program to explore the application of the model for calibrating operational data incorporating multiple items generated both from AIG and manual item generation. The sample consisted of 3,793 examinees in grade 8, distributed among 4 test forms. Results suggest that the inclusion of AIG generated items in item families will have a tendency to have item characteristic curves that are somewhat more variable than if the family consisted of the same item under repeated administration. However, this increased variability is neither assured, nor in most cases, even particularly pronounced. (Contains 2 figures, 1 table, and 13 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

The following paper was prepared for presentation at the March 2002 AERA/NCME program and is the copyrighted material of Educational Testing Service. Limited permission to reproduce reports in their entirety in downloadable PDF format is hereby granted for individual professional reference. Requests for additional permission to reproduce or use all or portions of these papers may be made through the ETS Permissions website by completing the convenient form at: http://www.ets.org/legal/copyright/html.

ED 464 949

Applying Hierarchical Model Calibration to Automatically Generated Items

David M. Williamson

Matthew S. Johnson

Sandip Sinharay

Isaac I. Bejar

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
D. Williamson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

Educational Testing Service

TM033880

BEST COPY AVAILABLE

Presented at the annual meeting of the American Educational Research Association
New Orleans, Louisiana
April, 2002

Applying Hierarchical Model Calibration to Automatically Generated Items

Recent research in educational measurement has been directed at methods to ensure an adequate and secure supply of items for item pools, particularly for continuous testing environments. Among these efforts are several lines of research targeted at the development of automatic item generation (AIG) systems; software capable of generating assessment items in a form requiring little or no human review prior to administration. These efforts are directed at various applications including verbal items (Sheehan, & Ginther, 2000), analytical reasoning (Dennis, Handley, Bradon, Evans, & Newstead, in press; Newstead, Bradon, Handley, Evans, & Dennis, in press), math (Singley, & Bennett, 2002), and abstract reasoning (Embretson, 1999). Of course, the extent to which items generated from these systems satisfy the needs of an assessment program depends on the purpose of the assessment and the particular item models developed and applied in the AIG software. Current efforts to develop AIG systems tend to have several elements in common, one of which is an emphasis on both cognitive and content modeling when developing operational item models. Another communality is an interest in the ability to predict item statistical performance from the item models used for AIG. Whether these AIG systems are eventually applied conjunctively with the efforts of human item writers or as the sole source of assessment items, these systems have the potential to substantially address the need for a large supply of items for operational item pools.

While such AIG systems, once implemented, would represent a substantial step toward providing items in abundance, the need for pretesting and calibration of these generated items would remain a bottleneck to operational use. Given that items

generated from a common item model may be expected to have a high degree of item dependence, there may be calibration models that would leverage this item dependence to facilitate a reduction in the need for pretesting these items for operational use. Ultimately, the successful development of AIG systems capable of producing items with highly similar statistical properties may permit the development and implementation of adaptive on-the-fly testing (Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, in press), in which an item pool does not actually exist and current ability estimates are used to generate items tailored for an examinee immediately prior to administration. This study explores the application of hierarchical model calibration as a means of reducing, if not eliminating, the need for pretesting of automatically generated items from a common item model prior to operational use.

Models for Related Items

While not unique to AIG, the inherent requirement of well-defined item models (also commonly called task models) in order to conduct automatic item generation facilitates the ability to exercise precise control over the degree of variation permitted in generated items. With the capability for such control, knowledge of the item model used in generation can provide information about the generative principles that produced the item. To the extent that item models used for AIG are based on research (e.g. cognition during task performance in the domain, domain-specific content principles, information processing research, etc.) the generated items have an underlying theoretical rationale for their use. This research base for the AIG item model can provide important evidence about *item pedigree*; documentation of the research foundation and history of design

decisions which spawned the model used for AIG and ultimately, the particular item in question.

On the basis of item pedigree, items can be assigned to an *item family*; a group of items believed to be closely related. (Exactly how closely related items must be to be considered family members can be defined by the user with respect to the information in the item pedigree and empirical evidence of item performance). In the case of AIG an obvious means for classification of item into item families is on the basis of the item model used for generation, with all items generated from a common model as members of a single item family. *Siblings* are items that are members of a common item family. Depending on the degree of control exercised in the item model used for generation, it can be expected that siblings would have a considerable degree of similarity in both content and statistical performance. Given that siblings share a common development rationale (through a common item model) and a corresponding expectation that their statistical performance will not be independent, there is fundamental question regarding the optimal way to model such related items in operational measurement.

Unrelated Siblings Model

The most conservative approach for calibration of item siblings is to treat the items as completely independent regardless of family membership. This unrelated siblings model is given by

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp\{a_i(\beta_i - \theta)\}} \qquad (1)$$

where *j* indicates the particular item in question. Since the model ignores the relationship between siblings in an item family the model is overly conservative, with use of these item response functions resulting in an unnecessarily large standard error for θ estimates.

Identical Siblings Model

A more liberal approach to calibration of item siblings is to consider siblings as having identical item response functions (Hombo & Dresher, 2001). This model is given by

$$P_j(\theta) = c_{I(j)} + \frac{1 - c_{I(j)}}{1 + \exp\left\{a_{I(j)}\left(\beta_{I(j)} - \theta\right)\right\}} \qquad (2)$$

where *I(j)* indicates the family of which item *j* is a member. Since the identical siblings model ignores all variation between siblings it results in inappropriately small standard errors for θ estimates, reflecting overconfidence about the ability of the examinee.

Related Siblings Model

A third alternative, utilized in the analyses for this paper, is to use a related siblings model in which each item is modeled with a separate item response function, but the siblings within a family are related by using a hierarchical model (Glas & van der Linden, 2001).

$$P_j(\theta_i) \equiv \Pr\{X_{ij} = 1 | \theta_i\} = c_j + \frac{1 - c_j}{1 + \exp\{a_j(\beta_j - \theta_i)\}} \quad (3)$$

where

$$\theta \sim N(\mu, \sigma^2)$$

$$\gamma_j \equiv \log\left\{\frac{c_j}{1 - c_j}\right\}$$

$$\alpha_j \equiv \log\{a_j\}$$

$$(\alpha_j, \beta_j, \gamma_j)' \sim N_3(\lambda_{I(j)}, T_{I(j)})$$

and where *i* indicates the examinee in question. This model appropriately accounts for sources of variation in responses: The responses of two individuals answering the same sibling are correlated. An additional advantage of this approach is that calibration of the item family and use of a family response function requires fewer observations for each item than calibration of each item individually.

This model is implemented in software (Johnson & Sinharay, 2002) that conducts Bayesian Markov Chain Monte Carlo (MCMC) estimation to estimate the joint posterior of all model parameters by integrating over the posterior distribution of model parameters given the data. The Monte Carlo integration draws samples from the required

distribution and then forms sample averages to approximate expectations. MCMC draws these samples by running a Markov chain for many iterations. As such, MCMC estimation is basically Monte Carlo integration using Markov chains; discrete time stochastic processes such that th. distribution of X_t (X at time t) depends only on X_{t-1} and is independent of all values X_{t-1} to X_{t-n}. Mathematically, this is represented as (Gilks, Richardson, & Spiegelhalter, 1996, p. 45):

$$P[X_t \in A \mid X_0, X_1, \ldots, X_{t-1}] = P[X_t \in A \mid X_{t-1}] \quad (4)$$

for any set A, where $P[.|.]$ denotes a conditional probability. For the related siblings model MCMC estimates the posterior distribution by drawing from the conditional posterior distribution of each model parameter. Item parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are drawn from their respective conditional distributions as described in Patz and Junker (1999). Conditional on the item parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the item family mean vector $\boldsymbol{\lambda}$ and the covariance matrix $\mathbf{T}$ are independent of θ and the observed data $\mathbf{X}$.

This study applies the related siblings model to math item data from an experimental administration associated with an ETS-operated national testing program in order to explore the application of the model for calibrating operational data incorporating multiple items generated both from AIG and manual item generation. We examine the similarity between item characteristic curves (ICCs)for the individual items and the item family response functions. If the family response functions are very similar to the individual sibling response functions then it may be appropriate to use the family response function as the AIG item model calibration, subsequently applying those parameters all items generated from that AIG item model (assuming proper model constraints) with little impact on θ estimates.

Method

Data

This study analyzed math item data from an experimental administration associated with an ETS-operated national testing program. The sample consisted of 3793 examinees in grade 8, distributed among four test forms. Each of the four forms had a block of common items (denoted MP) and an additional 26 mathematics items (denoted M2-M5 for the four forms), consisting of 16 multiple-choice and 10 open-ended items. The number of items of each type appearing in the four forms M2-M5 are presented in Table 1, as are the sample sizes from administration.

The 26 mathematics items comprising form M2 were written by human item writers and were assembled to be representative of the item pool, to the extent possible. This form was administered as a paper & pencil assessment, with one subset of items as a calculator-active block, with calculators provided for the students.

Form M3 is identical to form M2 and uses the same 26 items. However, this form was administered as a linear computerized assessment with an online calculator provided for the calculator-active block of items.

Form M4 was constructed to be parallel to form M2. Of the 26 items 11 were identical to the items appearing on form M2 while 15 items were automatically generated items (Singley, & Bennett, 2002) different from, but intended to be parallel to, the corresponding items on form M2. Like form M2, form M4 was administered via paper and pencil with a calculator provided for the calculator-active block.

Form M5 was constructed to be parallel to form M2. Of the 26 items 11 were identical to the items appearing on form M2 while 15 items were automatically generated items (Singley, & Bennett, 2002) different from, but intended to be parallel to, the corresponding items on form M2. The generated items for form M5, however, are different items than the generated items appearing on form M4. For each automatically generated item on form M4, there is a corresponding item generated from the same item model on form M5. Like form M2, form M5 was administered via paper and pencil with a calculator provided for the calculator-active block.

For this analysis the MP block was not considered and only the 16 dichotomously scored (multiple-choice) items of the other 26 items in each form were analyzed. In addition, there are no overlapping students in this design; that is, no one takes more than one of the forms.

Procedure

Data were analyzed with recently developed software (Johnson & Sinharay, 2002) that calibrates items using a hierarchical model (Glas & van der Linden, 2001) described above. The model applied prior distributions for the item family mean vectors that assumes the elements are independent and

$\lambda_a \sim N(0,100^2)$

$\lambda_b \sim N(0,100^2)$

$\lambda_c \sim N(-1.39,0.01)$

The prior density of the pseudo-guessing parameter (λ_c), when transformed to the c_j metric, has a mean at approximately .20 and a range from approximately .15 to .25.

The MCMC estimation procedure was conducted through 100,000 iterations, with the first 10,000 iterations treated as a burn-in period and therefore not included in the determination of the posterior distributions of the parameters. The remaining 90,000 iterations were thinned by selecting every 9th iteration for inclusion in the final data set determining the posterior distribution of the parameters. This resulted in a final data set consisting of 10,000 draws for the distribution of each parameter. The item characteristic curves (ICC) were produced using the median value of the distribution for each parameter. The root-mean-square-error (RMSE) was computed for the ICCs for each family, using the family calibration as the ICC for comparison of the item ICCs in the computation. The RMSE is given by

$$RMSE = \sqrt{\frac{\sum_{t=-3.0}^{3.0} (p_{it} - p_{ft})^2}{n_t}} \quad , \qquad (5)$$

where p_{it} indicates the item ICC probability of responding correctly at ability t, p_{ft} indicates the family ICC probability of responding correctly at ability t, and n_t is the number of theta values considered (in this case using the values between –3.0 and 3.0 in intervals of .1, so n_t=61).

Results

The ICCs and family characteristic curves are provided by family as Figure 1, with the families without any AIG items preceding those containing AIG items (indicated by a parenthetical AIG after the family identifier). Those item families without AIG items generally have more closely corresponding ICCs than families with AIG items, with the most similar set of ICCs represented in family 52301. This is, of course, not

surprising considering the fact that families without AIG items are presenting a series of ICCs all on the same item appearing in different forms. Despite the generally close ICCs for item families without AIG items there is some variation evident in some of the ICCs for these families, with the greatest observed variation evident for family 18301.

Examination of the families that contain AIG items reveals a couple of immediately obvious deviations. Most obvious is the fact that the entire family of items for family 52801 is flat at approximately random chance for all levels of ability. Since this is true for both the human generated item (appearing in form M2 and M3) and the AIG items (appearing in form M4 and M5) and the ICCs are consistent with the classical statistics calculated on the items it would appear that this is the result of a characteristic of the item type or content rather than the result of anything inherent in AIG.

Another obvious deviation in ICCs occurs in family 72801. In this instance the manually generated item and the AIG item appearing in M5 have very similar ICCs while the AIG item appearing in M4 deviates dramatically from the other items in the family. The extent of the deviation also appears to impact the response function for the family as a whole.

In the case of family 51401 the correspondence between the ICCs for the human generated item and the AIG items is close but there is an obvious difference in the pseudo-guessing parameter between the item ICCs and the family response function that appears to be an artifact of the range of prior selected for the parameter. Some of the other families also have fairly minor deviations of the ICC for one of the AIG items from the ICCs for the others, including family 67401, 67301, 11131, and 13731.

A number of the families with AIG items appear to have ICCs that are quite similar for both the human generated item and the AIG items. These include families 46301, 12431, 13531, and 73301. Still others, including families 12431 and 73301, have ICCs for the AIG items that are as close or closer to the ICC for one administration of the human generated item than even the ICC for the other administration of the same human generated item is.

The plot of the RMSE for the families without AIG items and the families that included AIG items are provided as Figure 2(a) and Figure 2(b), respectively. Examination of Figures 2(a) and 2(b) further suggests that there are generally lower RMSE for item families that do not have AIG than for the families that do have AIG items. For those item families that incorporate AIG items it would appear that the ICCs for the human produced items are generally about the same approximation to the family response function as for the AIG item ICCs. Of course, when considering this result one must remain aware that for families with AIG items the family response function was generated for the entire family, which includes an equal mixture of two AIG items and two human generated items.

Discussion

These results suggest that the inclusion of AIG generated items in item families will have a tendency to have ICCs that are somewhat more variable than if the family consisted of the same item under repeated administrations. However, this increased variability is neither assured nor in most cases even particularly pronounced. While some item families demonstrated some variability in ICCs as a result of one of the AIG

generated items, many others were very similar and approximated the ICC consistency observed in families that used the same item repeatedly on each form.

With the exception of a single of notable case (family 72801) the range of RMSEs computed from the AIG item families are similar to the range of RMSEs obtained from a study (Rizavi, Way, Davey, & Herbert, April, 2002) in which the same subset of items from Verbal and Quantitative sections of a high-stakes admissions test were recalibrated through eight administrations and the variation in item parameters evaluated. If variations in ICCs for item families that use AIG generated items tasks are consistently similar to variations obtained from recalibration of the same multiple-choice item over repeated administrations then there is some evidence that the AIG item models can be leveraged to produce multiple parallel items that have highly similar statistical properties

Despite the apparent degree of similarity from the calibration of AIG generated items in the item families a number of important research issues remain outstanding before fully committing to the operational application of family response functions to all items in a family. Specifically, it will be important to establish the degree of variation in θ estimates as a result of the observed variation in ICCs among siblings that include a wider range of AIG items. Furthermore, as a result of potential parameter variation it will be important to establish the possible implications on ability estimates and subsequent decision making (e.g. placement decisions, licensure, etc.) in operational environments.

Researchers in the field have recognized the importance of these issues and have already begun to address them. Dresher & Hombo (2001), for example, investigated the impact of simulated parameter variation on ability estimation and concluded that ability

estimation, for both individuals and grouped score reporting, was largely robust to variation in parameter estimates. Similar conclusions regarding the feasibility of operational use of AIG items were reached in a related investigation of item parameter bias in simulated NAEP-like assessment conditions (Hombo & Dresher, 2001). The impact of AIG item parameter variation on ability estimates has also been addressed by by Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta (in press) for on-the-fly adaptive testing . As a result of these investigations and other ongoing research a full perspective on operational application of AIG items using a common family parameterization is becoming more fully developed and may be paving the way for the eventual operational use of AIG items with common family parameterizations, furthering the potential for adaptive on-the-fly assessment.

References

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (in press). *A feasibility study of on-the-fly adaptive testing* (RR-xx-xx). Princeton, NJ: Educational Testing Service.

Dennis, I., Handley, S., Bradon, P., Evans, J., & Newstead, S.E. (in press). Approaches to modeling item generative tests. In Kyllonen, P. and Irvine, S.H. (Eds.) *Item Generative Testing*.

Dresher A, & Hombo, C. (2001). *A simulation study of the impact of automatic item generation on item and ability parameter estimation*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika, 64*, 407-433.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (1996). *Markov chain monte carlo in practice*. New York: Chapman & Hall.

Glas, C. A. W. & van der Linden, W. J. (2001). *Modeling variability in item parameters in CAT*. Paper presented at the North American Psychometric Society Meeting, King of Prussia, PA.

Hombo, C. & Dresher, A. (2001). *A simulation study of the impact of automatic item generation under NAEP-like data conditions*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Johnson, M. S. & Sinharay, S. (April, 2002). *Hierarchical approaches to item model calibration.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Newstead, S.E., Bradon, P., Handley, S., Evans, J., and Dennis, I. (in press). Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In Kyllonen, P. and Irvine, S.H. (Eds.) *Item Generative Testing.*

Patz, R. & Junker, B. (1999). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

Rizavi, S., Way, W. D., Davey, T., & Herbert, E. (April, 2002) *Tolerable variation in item parameter estimation.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Sheehan, K. M., & Ginther, A. (2000). *What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development.* Hillsdale, NJ: Erlbaum.

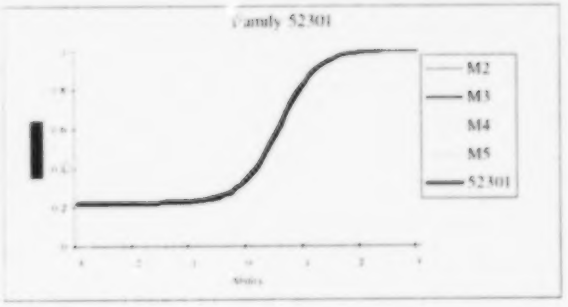
Table 1

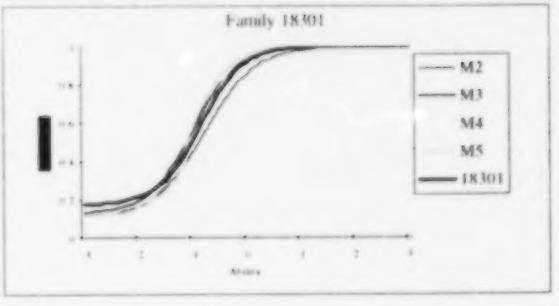
Item Type and Generation by Form

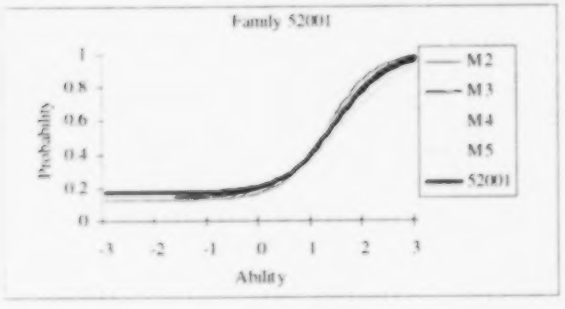
	Graded Response		Multiple Choice		
Form	Human	AIG	Human	AIG	Sample Size
M2	10		16		1014
M3	10		16		953
M4	6	4	5	11	922
M5	6	4	5	11	904

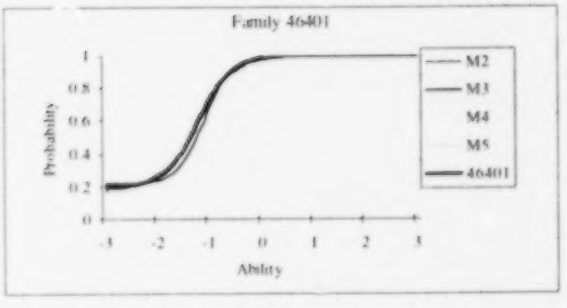
Figure 1

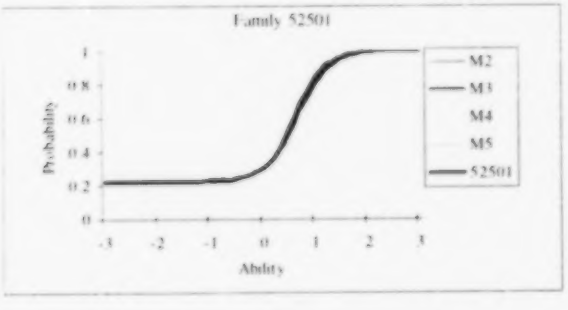
Item and Family Characteristic Curves



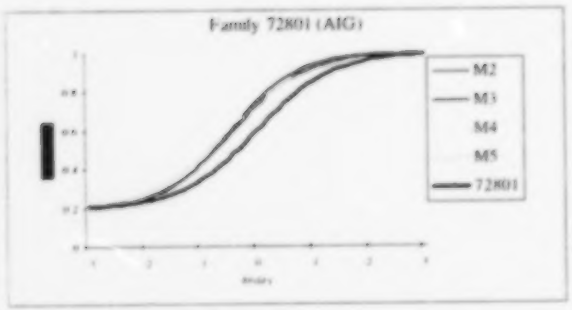


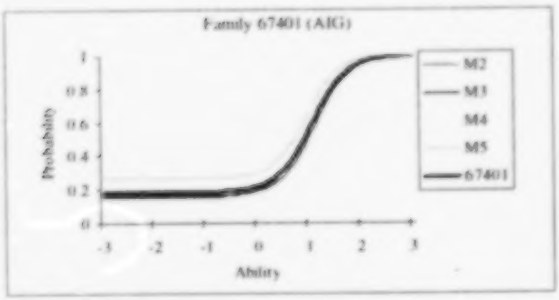


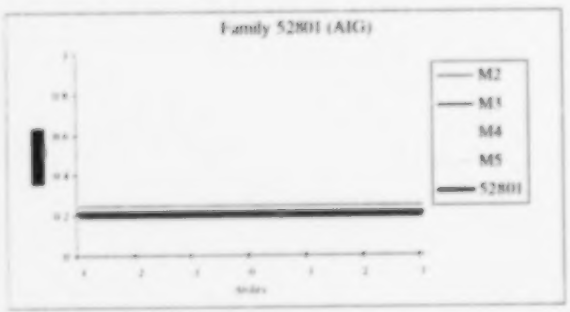


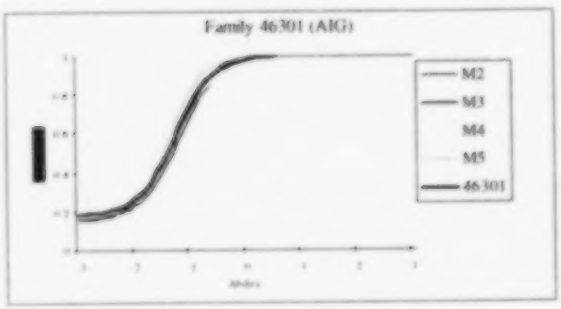


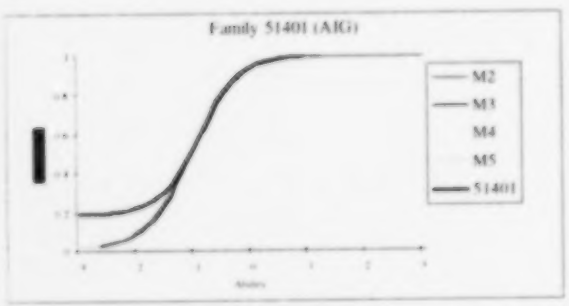
BEST COPY AVAILABLE

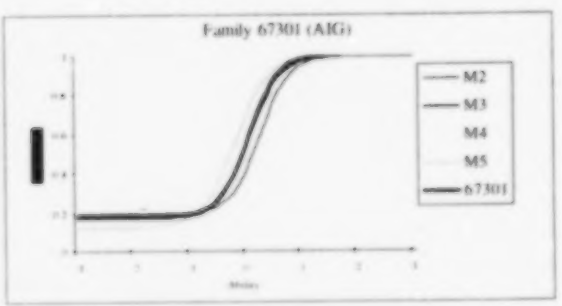

Family 72801 (AIG)
M2
M3
M4
M5
72801

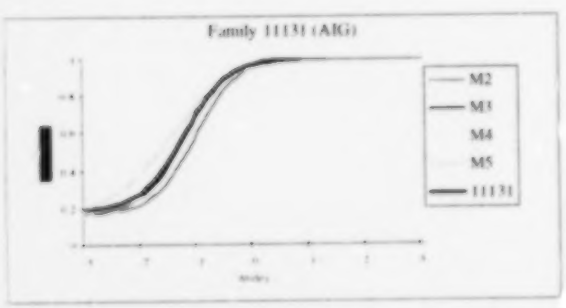

Family 67401 (AIG)
Probability
Ability
M2
M3
M4
M5
67401

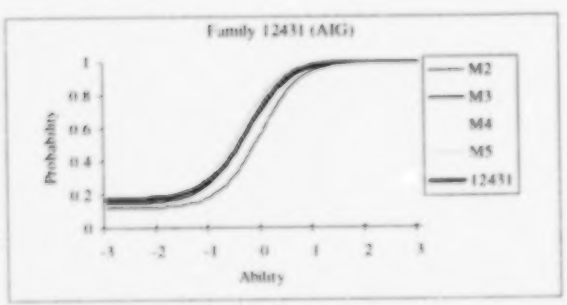

Family 52801 (AIG)
M2
M3
M4
M5
52801

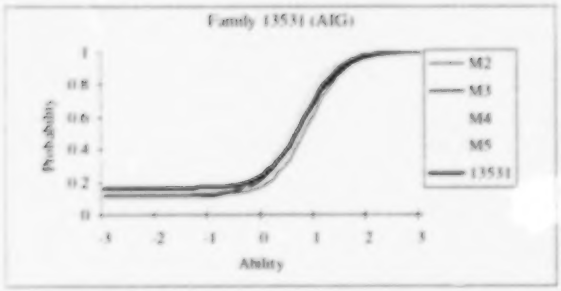

Family 46301 (AIG)
M2
M3
M4
M5
46301

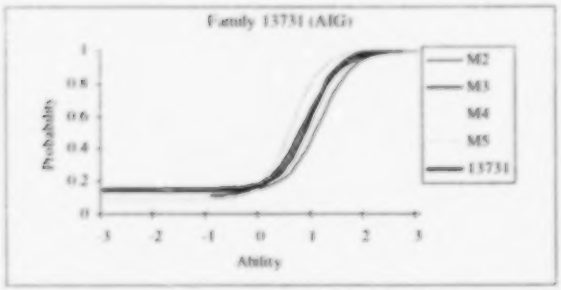

Family 51401 (AIG)
M2
M3
M4
M5
51401


Family 67301 (AIG)
M2
M3
M4
M5
67301


Family 11131 (AIG)
M2
M3
M4
M5
11131


Family 12431 (AIG)
Probability
Ability
M2
M3
M4
M5
12431


Family 13531 (AIG)
Probability
Ability
M2
M3
M4
M5
13531


Family 13731 (AIG)
Probability
Ability
M2
M3
M4
M5
13731

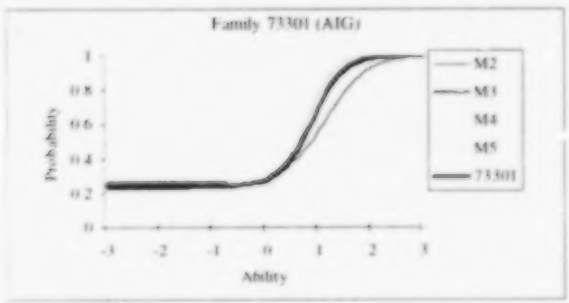

Family 73301 (AIG)
Probability
Ability
M2
M3
M4
M5
73301

Figure 2(a)

Root Mean Squared Error for Families Without AIG Items

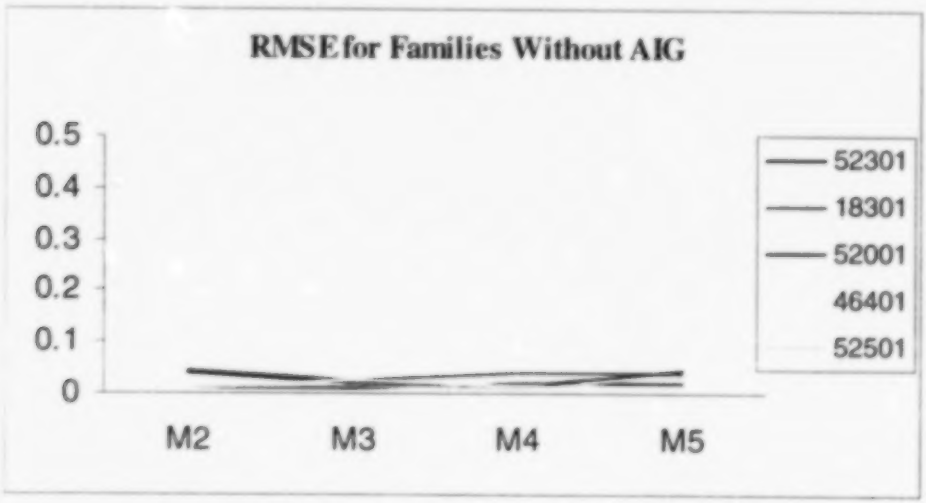


Figure 2(b)

Root Mean Squared Error for Families With AIG Items